

When the Chain Breaks: Interactive Diagnosis of LLM Chain-of-Thought Reasoning Errors



Shiwei Chen¹, Niruthikka Sritharan¹, Xiaolin Wen¹, Chenxi Zhang¹, Xingbo Wang², Yong Wang¹

1 – Nanyang Technological University, Singapore



2 – Bosch Research North America, Sunnyvale, California, USA



EuroVis 2026

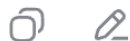
The Eurographics Conference on Visualization



The Problem

What is the gap between EuroVis 2025 and EuroVis 2026, held by Eurographics?

Thinking 



- We need to find the gap between EuroVis 2025 and EuroVis 2026, held by Eurographics. The user is asking about the time gap between the two conferences. I need to recall or find information about these conferences. EuroVis is the Eurographics Conference on Visualization, usually held annually. The gap likely refers to the time interval between the two events, perhaps in days or months. I need to determine the dates of EuroVis 2025 and EuroVis 2026.

I know that EuroVis 2025 is scheduled for June 2025 in Luxembourg? Let's check. Actually, EuroVis 2025: I recall the EuroVis 2025 website: "EuroVis 2025 will take place in Luxembourg, June 9-13, 2025." Wait, I need to confirm. I think EuroVis 2025 is June 9-13, 2025, in Luxembourg City. EuroVis 2026: I need the dates announced. I recall that EuroVis 2026 is planned for June 8-12, 2026 in Porto, Portugal. Let's search my memory. EuroVis 2026: The Eurographics conference is held annually. EuroVis 2026: June. I remember that EuroVis 2026 will be held in Porto? Or Lisbon? I think it's Porto, Portugal. Actually, the EuroVis 2024 was in Odense, Denmark, and EuroVis 2025 is in Luxembourg, June 9-13, 2025. EuroVis 2026: I think I've seen "EuroVis 2026 will be held in Porto, Portugal, June 8-12, 2026" or "Lisbon". Let me think: In EuroVis 2025 website, they often announce the next year's

87 steps. one question.

find the "one" wrong step?

The Problem

What is the gap between EuroVis 2025 and EuroVis 2026, held by Eurographics?

Thinking ▾

- We need to find the gap between EuroVis 2025 and EuroVis 2026, held by Eurographics. The user is asking about the time gap between the two conferences. I need to recall or find information about these conferences. EuroVis is the Eurographics Conference on Visualization, usually held annually. The gap likely refers to the time interval between the two events, perhaps in days or months. I need to determine the dates of EuroVis 2025 and EuroVis 2026.

I know that EuroVis 2025 is scheduled for June 2025 in Luxembourg? Let's check. Actually, EuroVis 2025: I recall the EuroVis 2025 website: "EuroVis 2025 will take place in Luxembourg, June 9-13, 2025." Wait, I need to confirm. I think EuroVis 2025 is June 9-13, 2025, in Luxembourg City. Let me verify: EuroVis 2025 dates: June 9 to 13, 2025. Yes, I've seen that. EuroVis 2026: I need the dates for 2026. EuroVis 2026 might be announced. I recall that EuroVis 2026 is planned for June 8-12, 2026 in Lisbon, Portugal? Wait, I'm not sure. Let's search my memory. EuroVis 2026: The Eurographics conference schedule usually has EuroVis in early June. I remember that EuroVis 2026 will be held in Porto? Or Lisbon? I recall something about EuroVis 2026 in Porto, Portugal. Actually, the EuroVis 2024 was in Odense, Denmark, May 27-31, 2024. EuroVis 2025: Luxembourg, June 9-13, 2025. EuroVis 2026: I think I've seen "EuroVis 2026 will be held in Porto, Portugal, June 8-12, 2026" or "Lisbon". Let me think: In EuroVis 2025 website, they often announce the next year's venue. I can recall the closing session of EuroVis 2024 announced EuroVis 2025 in Luxembourg, and EuroVis 2026 in Porto. Yes, I think I saw that: "EuroVis 2026 will be held in Porto, Portugal, from June 8 to 12, 2026. Let's confirm: I remember a tweet or post: "EuroVis 2026 will be in Porto, June 8-12, 2026." So, EuroVis 2025: June 9-13, 2025; EuroVis 2026: June 8-12, 2026. Then the gap between the two: The end of EuroVis 2025 is June 13, 2025, and the start of EuroVis 2026 is June 8, 2026. The gap in days would be from June 14, 2025 to June 7, 2026, which is roughly 358 days? Let's calculate exactly. From June 13, 2025 to June 8, 2026. Number of days between these dates inclusive or exclusive? Usually "gap between" might mean the time interval between the end of one and the start of the next. Often conferences are annual, so gap is about a year. I can answer: EuroVis 2025: 9-13 June 2025; EuroVis 2026: 8-12 June 2026 (expected). The gap is approximately one year, specifically the interval from the closing of EuroVis 2025 (13 June 2025) to the opening of EuroVis 2026 (8 June 2026) is 360 days (or 11 months 26 days). Need to compute exactly.

Let's compute days: From June 13, 2025 to June 8, 2026.

- June 14, 2025 to June 8, 2026. Count days in between. Use date difference: June 8, 2026 minus June 13, 2025. June 13, 2025 to June 13, 2026 is 365 days (2026 is not a leap year? 2026 is not divisible by 4, so 365 days). Then subtract 5 days to get to June 8, 2026: $365 - 5 = 360$ days. So the gap between the last day of EuroVis 2025 (June 13) and the first day of EuroVis 2026 (June 8) is 360 days. But if we consider the whole conference dates, the midpoint or something. The user asked "What is the gap between EuroVis 2025 and EuroVis 2026" likely meaning the time interval between the two conferences. I'll answer with exact dates if I have them, and compute the gap. I'll mention that EuroVis 2026 is scheduled for June 8-12, 2026 in Porto, Portugal, while EuroVis 2025 is June 9-13, 2025 in Luxembourg. The gap from the end of one to the start of the next is 360 days (just under 1 year). Provide both start-to-start gap: June 9, 2025 to June 8, 2026 is 364 days? June 9, 2025 to June 9, 2026 is 365 days, then minus 1 day = 364 days. So start-to-start is 364 days (since one year non-leap minus 1 day). I'll note that the conference is annual with a typical one-year gap.

Make sure to verify EuroVis 2026 location and dates. I'm fairly confident it's Porto, Portugal, June 8-12, 2026. I'll double-check by reasoning: EuroVis 2024 was Odense, May 27-31. EuroVis 2025 Luxembourg, June 9-13. EuroVis 2026: I recall a slide from EuroVis 2024 closing session: "EuroVis 2026 will be in Porto, Portugal, date: June 8-12, 2026." Yes, I remember that announcement. I can also check recent Eurographics events: EG 2025 in London, EG 2026 in ... but EuroVis is separate. The EG conference will be in Rennes? Not relevant. I'll structure answer: state dates and venues, compute gap. I'll mention that as per current scheduling, the gap is about 360 days between conferences, with specific dates. I'll also note that the exact gap might be measured in different ways: from the end of EuroVis 2025 to the start of EuroVis 2026 is 360 days, from the start to start is 364 days. Provide that.

Thus answer.

✓ Explore many possible paths

● Verbosity:

- Easily hundreds of steps
- Complex logic

✗ Hard to build a coherent mental model

✓ Explicitly state reasoning steps

● Unreliability:

- Contain subtle mistakes
- Answer overconfidently

✗ Cause misplaced trust



Research Gap

Provider	Model (example)	API reasoning exposure	UI reasoning exposure
OpenAI [Ope26]	GPT-5	summary	summary
Google [Goo26]	Gemini 3	summary	summary
Anthropic [Ant26]	Claude Opus 4	raw reasoning	summary
xAI [xAI26]	Grok 3/4	no	summary
DeepSeek [Dee26]	DeepSeek-R1	raw reasoning	raw reasoning
Qwen [Qwe26]	Qwen3/3.5	raw reasoning	raw reasoning
Mistral [AI26]	magistral	raw reasoning	raw reasoning

Table 1: A summary of 7 LRM providers in terms of the reasoning trace exposure in their user interfaces and APIs, as of Mar. 2026.

- Hidden by default

Thinking ∨ 

- Click and show summary

Explaining Mother's Day date gap variations >

The user wants to know if the gap between Mother's Day is always 364 days.

It shows only high level description and loses context.

It risks overconfidence and over-reliance.

Formative Study

Process

9 participants with intense LLM usage

Each was given an long CoT sample from the *Deltabench*;

What is their practice?
What is the main difficulty?

Findings

F1 Users prioritize **different steps**.

F2 Long CoTs cause **cognitive overload**.

F3 **Nonlinear logic** kills traceability.

F4 **Limited verification** resources.

derive



Design requirements

R1 **Automatically detect** potential errors.

R2 Provide an **overview** of the overall reasoning.

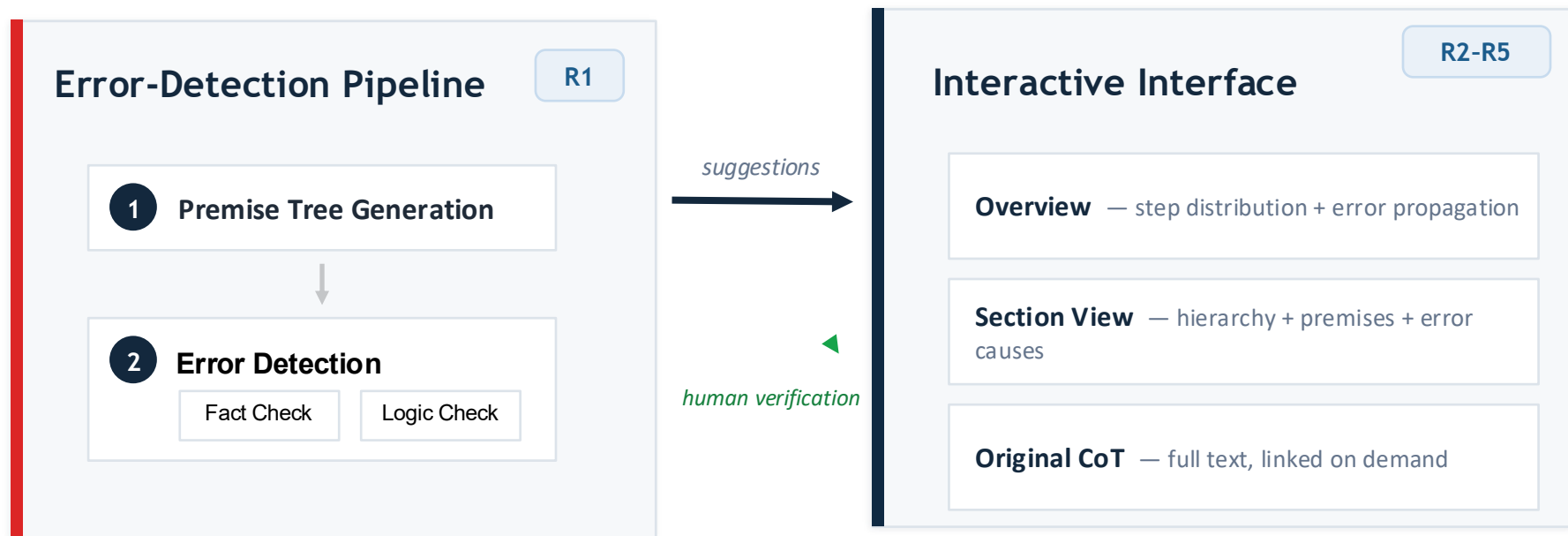
R3 Support traceable **navigation** across steps.

R4 Visualize how they **propagate**.

R5 Enable **cause** analysis of individual errors.

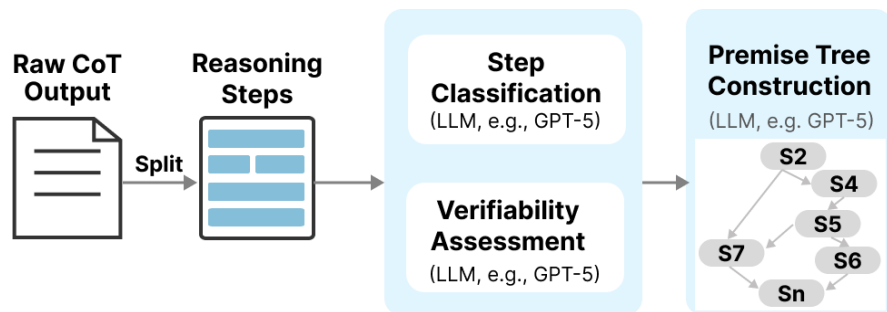
System Overview

ReasonDiag = automatic detection + interactive diagnosis

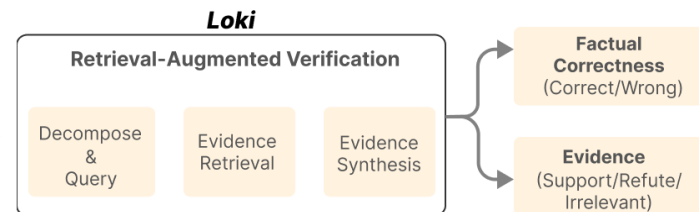


Error Detection Pipeline

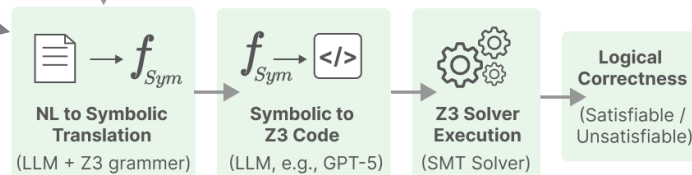
Step 1. Premise Tree Generation



Step 2. Error Detection

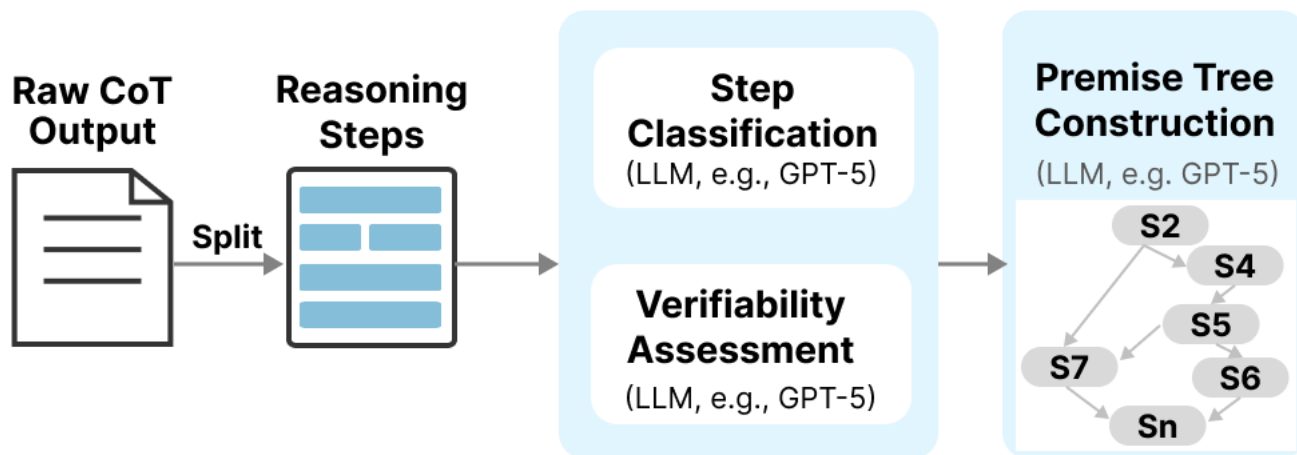


Factual Error Detection

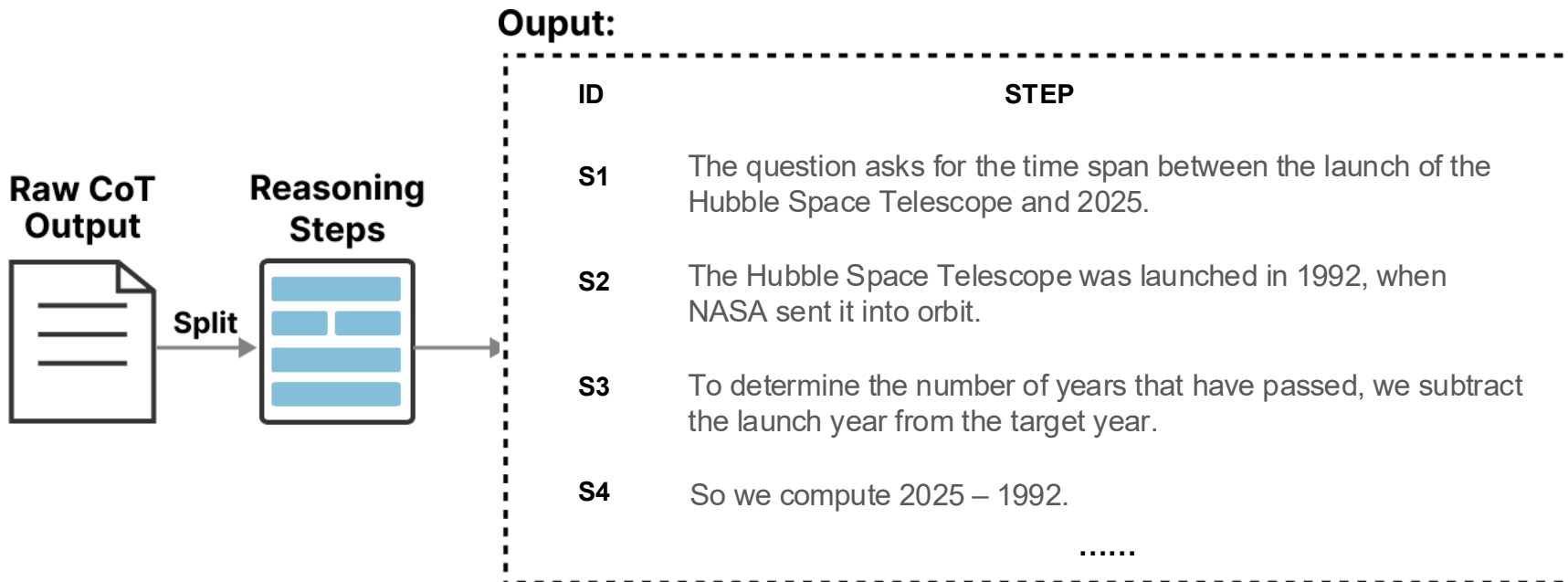


Logical Error Detection

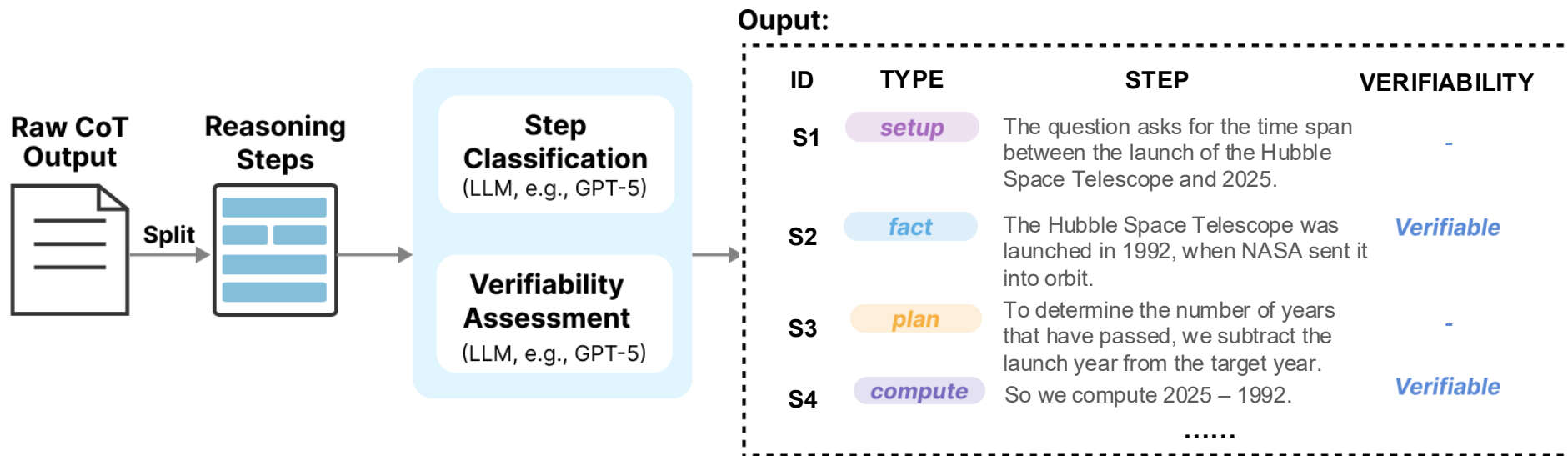
Step 1. Premise Tree Generation



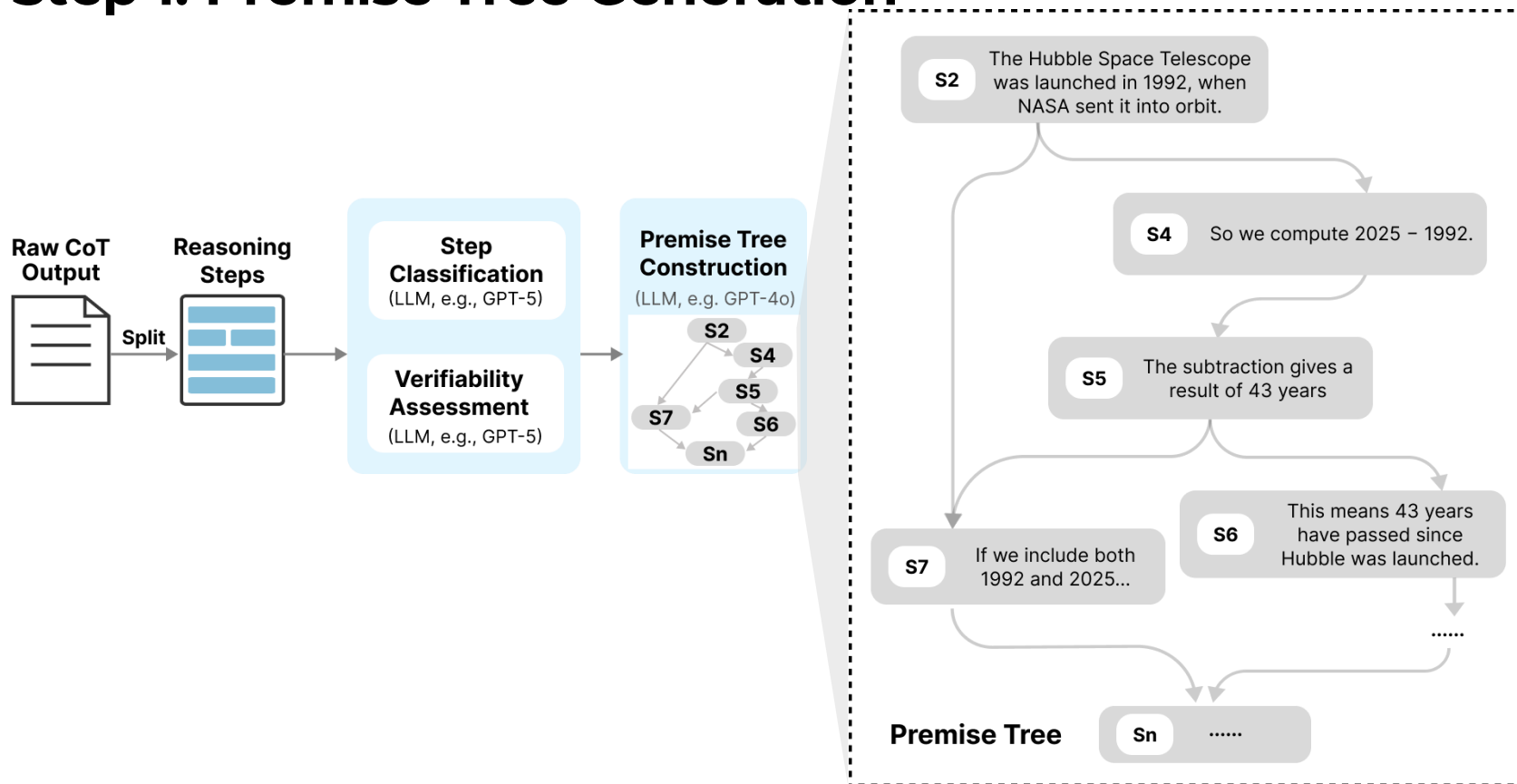
Step 1. Premise Tree Generation



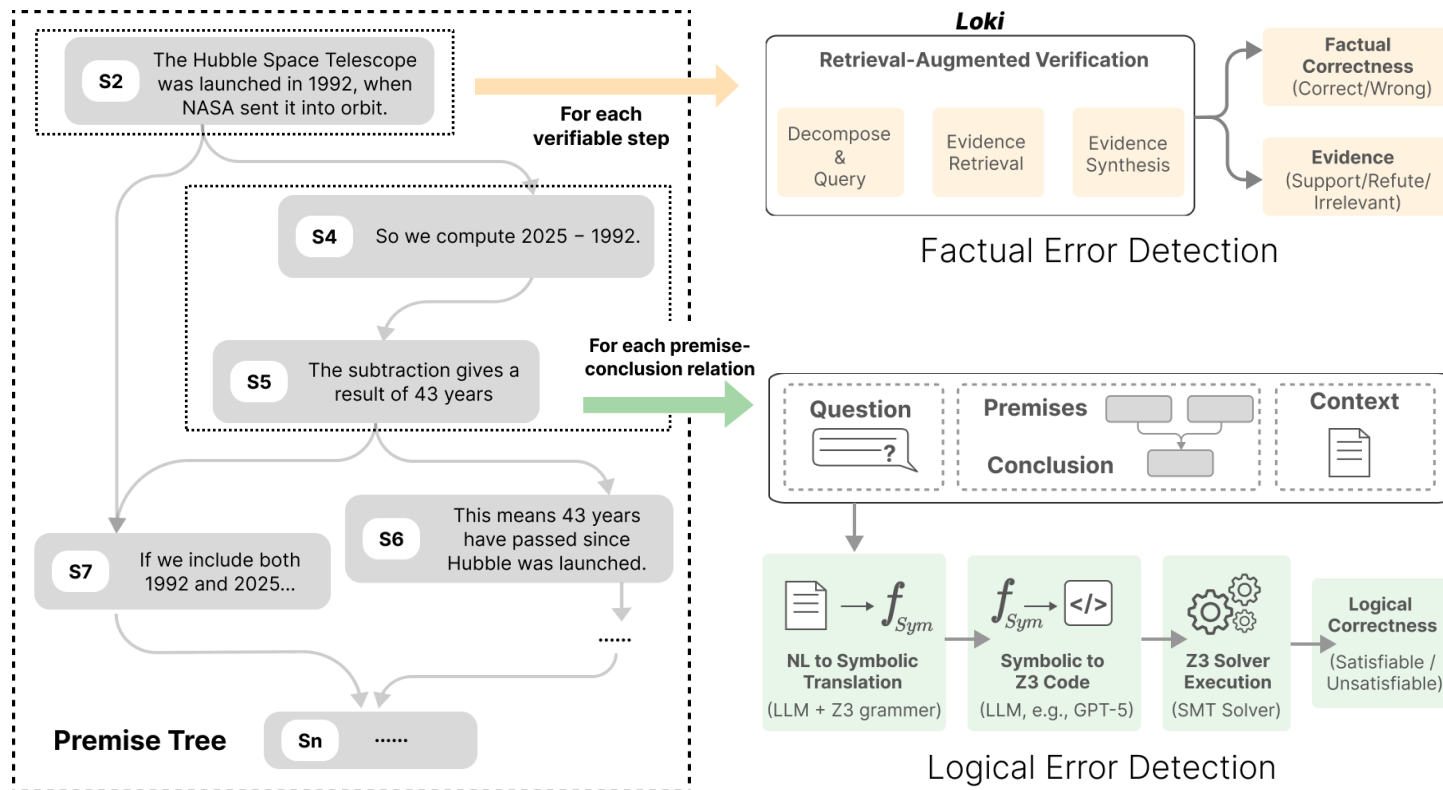
Step 1. Premise Tree Generation



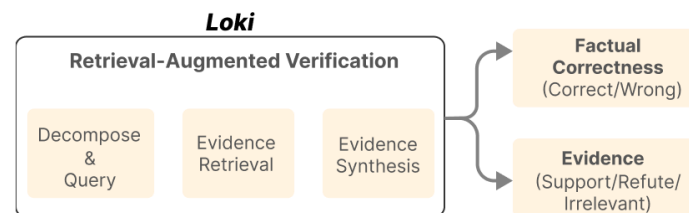
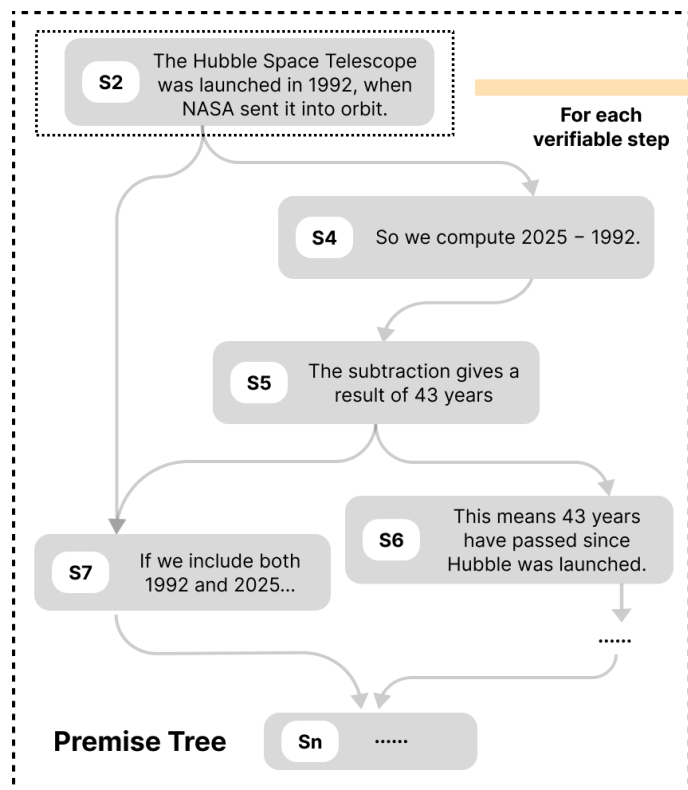
Step 1. Premise Tree Generation



Step 2. Error Detection



Step 2.1 Factual Error



Factual Error Detection

FactError [Close]

Sentence

The Hubble Space Telescope was launched in 1992, when NASA sent it into orbit.

Reason

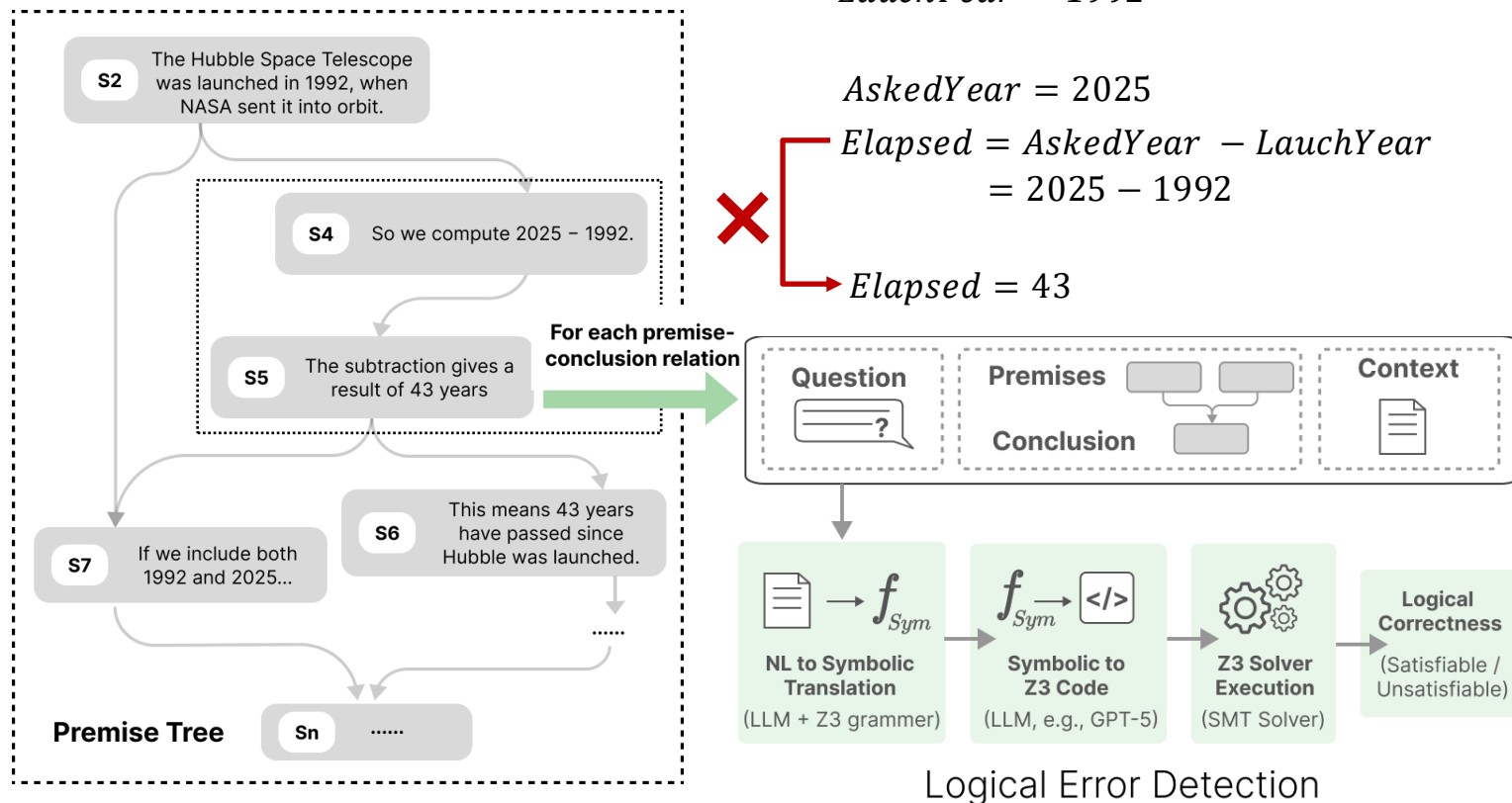
Fact Check Evidence [All] **Refutes** [Supports]

REFUTES [Source]

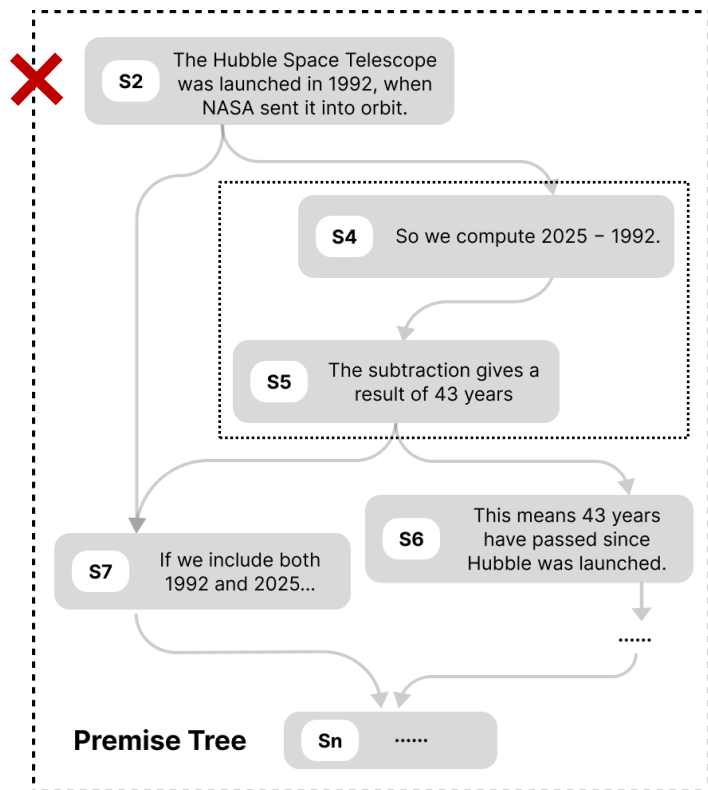
"The Hubble Space Telescope was launched into low Earth orbit on April 24, 1990, aboard the Space Shuttle Discovery (STS-31)."

The evidence explicitly states that Hubble was launched in April 1990, which contradicts the statement that it was launched in 1992.

Step 2.2 Logical Error



Step 2.3 Propagated Error



FactError ✕

Sentence

The Hubble Space Telescope was launched in 1992, when NASA sent it into orbit.

Reason

All **Refutes** Supports

Fact Check Evidence

REFUTES Source ↗

"The Hubble Space Telescope was launched into low Earth orbit on April 24, 1990, aboard the Space Shuttle Discovery (STS-31)."

The evidence explicitly states that Hubble was launched in April 1990, which contradicts the statement that it was launched in 1992.

X

$LauchYear = 1992$

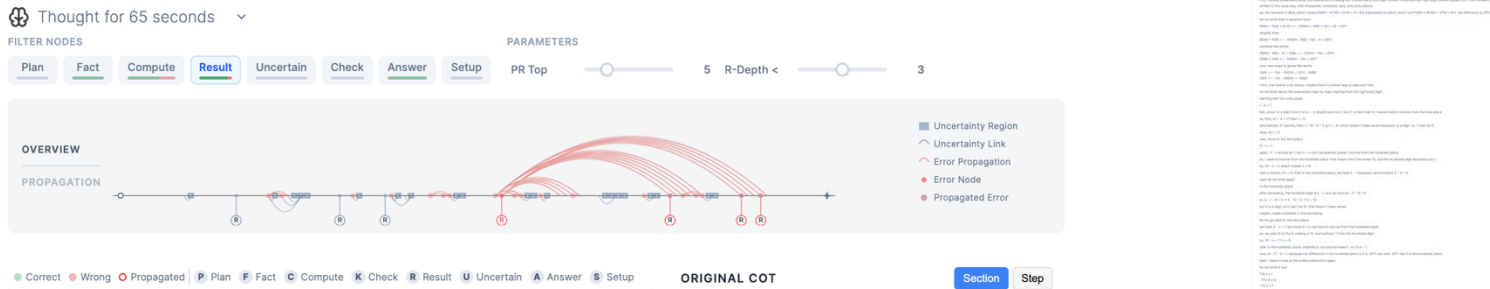
$AskedYear = 2025$

$Elapsed = AskedYear - LauchYear$
 $= 2025 - 1992$

$Elapsed = 33$

ReasonDiag UI

Overview



Section View



ORIGINAL COT

Section Step

setup So i've got this subtraction problem here:

??6 k 0 l
 -??m 9 n 4
 =22011

and i need to find the digits k, l, m, and n, and then add them up to find k + l + m + n. **setup** first, i should understand what this subtraction is telling me. **setup** it looks like a four-digit number minus another four-digit number equals 2011. **setup** the numbers are written in the usual way, with thousands, hundreds, tens, and units places.

fact so, the minuend is 6k0l, which means $6000 + k \cdot 100 + 0 \cdot 10 + l \cdot 1$. **fact** the subtrahend is m9n4, which is $m \cdot 1000 + 9 \cdot 100 + n \cdot 10 + 4 \cdot 1$. **fact** the difference is 2011. **compute** let me write that in equation form:

$$6000 + 100k + 0 \cdot 10 + l - (1000m + 900 + 10n + 4) = 2011$$

simplify that:

$$6000 + 100k + l - 1000m - 900 - 10n - 4 = 2011$$

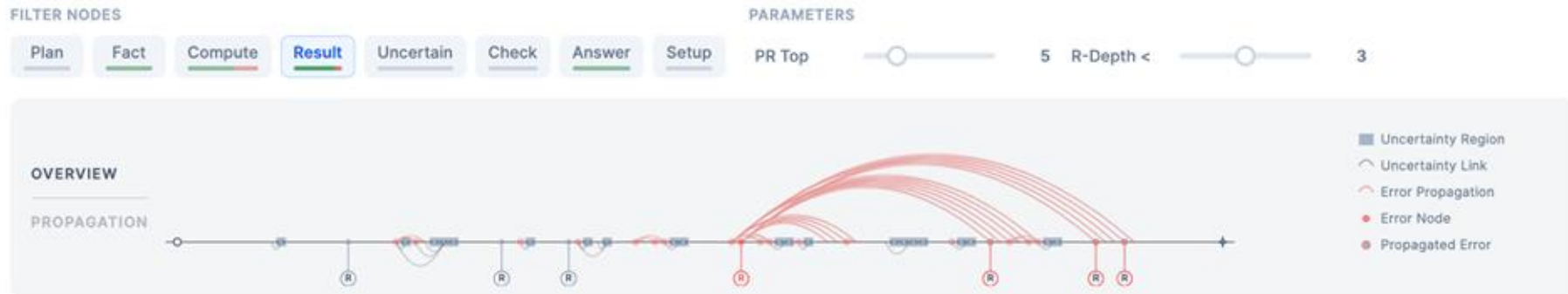
combine like terms:

$$(6000 - 900 - 4) + 100k + l - 1000m - 10n = 2011$$

$$5096 + 100k + l - 1000m - 10n = 2011$$



Overview



	Node	Link
Error	<ul style="list-style-type: none"> ● an erroneous step ● an propagated error step 	<ul style="list-style-type: none"> ⤴ error source --- propagated error
Uncertainty	<ul style="list-style-type: none"> ■ uncertain step 	<ul style="list-style-type: none"> ⤴ “uncertained” step --- uncertain step

Retroactive reasoning patterns



compute wait a minute, if $n = 9$, then in the hundreds place, we have $k - 1$ (because we borrowed 1) - 9 = 0.

uncertain wait, let me think again.

compute in the hundreds place:

after borrowing, the hundreds digit is $k - 1$, and we have $(k - 1) - 9 = 0$.

compute so, $k - 1 - 9 = 0 \rightarrow k - 10 = 0 \rightarrow k = 10$.

uncertain but k is a digit, so it can't be 10.

uncertain that doesn't make sense.

uncertain maybe i made a mistake in the borrowing.

In the subtraction shown, K , L , M , and N are digits. What is the value of $K + L + M + N$?

$$\begin{array}{r} 6K0L \\ - M9N4 \\ \hline 2011 \end{array}$$

compute $n = 9$

hundreds: $(k - 1) - 9 = 0$?

compute $k - 10 = 0$?

uncertain $k = 10$, which is impossible.

uncertain wait, that can't be right because k has to be a single digit.

Section View

PARAMETERS

PR Top

5

R-Depth <

3

● Correct ● Wrong ○ Propagated | P Plan F Fact C Compute K Check R Result U Uncertain A Answer S Setup

Problem Understanding

| S S S S

Formulating Equation

| F F F C U

Digit-by-Digit Subtraction

| P

Resolving Units Place

| C C C C C R

Resolving Tens Place

| C C C C C U

Resolving Hundreds Place

| C C U U U

Correcting Tens Place

| P C C C R C C U

Revisiting Subtraction

| P C C R C U F U

Rechecking Units and Tens

| P C C C C C U U

Resolving Thousands and Hundreds

| P P C C C R P C C U U C U

Step-by-Step Subtraction

| P C C C C P C C U U U P C C U U

Finding Final Digit

| P R C C C C C U U P C C

Summing the Digits

| R

Verifying the Solution

| K K R A

LogicError

SENTENCE
m = 4 okay, so m = 4.

REASON
Dependency Graph

so, in the hundreds place, if $k - 1 - 9$ is not enough, maybe I need to borrow from the thousands place.

so, the thousands digit is 6, and m is being subtracted.

let me consider the thousands place: $6 - m = 2?$

$m = 4$ okay, so $m = 4$.

In the subtraction shown, K , L , M , and N are digits. What is the value of $K + L + M + N$?

$$\begin{array}{r} 6K0L \\ - M9N4 \\ \hline 2011 \end{array}$$

compute

so, in the hundreds place, if $k - 1 - 9$ is not enough, maybe I need to borrow from the thousands place.

compute

so, the thousands digit is 6, and m is being subtracted.

compute

let me consider the thousands place:

$$6 - m = 2?$$

result

$$m = 4$$

okay, so $m = 4$.

Technical Evaluation

Dataset:

13 CoT samples from *Deltabench*
2,030 sentence-level annotations

Method	Precision.	Recall.	F1
GPT-5 (BIG-Bench Prompt)	0.432	0.658	0.470
GPT-5 (DeltaBench Prompt)	0.051	0.041	0.044
<i>ReasonDiag (Ours)</i>	0.306	0.801	0.386

Highest Recall

Covering the most errors

User Interviews

16 LLM users with diverse background (aged 18-31)

Workflow Effectiveness:

Question	Strongly Disagree	Mostly Disagree	Disagree	Neutral	Agree	Mostly Agree	Strongly Agree	Mean	Std
Q1. With ReasonDiag, I can easily understand the reasoning trace.					5	7	4	5.94	0.77
Q2. With ReasonDiag, I can easily judge if the reasoning trace is correct.					2	9	4	6.00	0.82
Q3. With ReasonDiag, I can easily locate erroneous steps.					1	10	5	6.25	0.58
Q4. With ReasonDiag, I can easily determine error causes.	1				2	11	2	5.81	0.91
Q5. Overall, ReasonDiag helps me diagnose errors in CoT reasoning trace effectively.					4	7	5	6.06	0.77

Design and Interactions:

Question	Strongly Disagree	Mostly Disagree	Disagree	Neutral	Agree	Mostly Agree	Strongly Agree	Mean	Std
Q6. The Overview clearly conveys error propagation patterns across the reasoning trace.	1				6	5	3	5.44	1.26
Q7. The Overview helps me understand type distribution of reasoning steps.					2	6	7	6.19	0.91
Q8. The Section View helps me understand the reasoning process at the section level.					1	9	5	6.13	0.81
Q9. The Section View clearly shows premise-conclusion relations of reasoning steps.	1				1	7	5	5.81	1.22

Usability:

Question	Strongly Disagree	Mostly Disagree	Disagree	Neutral	Agree	Mostly Agree	Strongly Agree	Mean	Std
Q10. I thought ReasonDiag is easy to learn.					6	8	2	5.75	0.68
Q11. I thought ReasonDiag is easy to use.					5	7	4	5.94	0.77
Q12. I would like to use ReasonDiag frequently.	1				3	5	5	5.69	1.25
Q13. I would like to recommend ReasonDiag to others.	1				1	8	5	5.94	1.12

Question	Strongly Disagree	Mostly Disagree	Disagree	Neutral	Agree	Mostly Agree	Strongly Agree	Mean	Std
Q14. What are the pros and cons of ReasonDiag?									
Q15. What improvements would you suggest for ReasonDiag?									

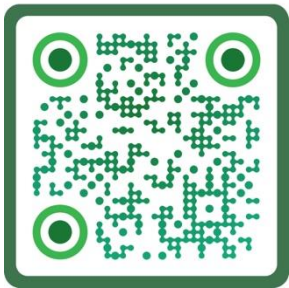
Effective,
Well-designed,
Good usability

Takeaway

- CoT reasoning traces generated by LLMs can be problematic
- ReasonDiag is an attempt to use visualization help general users easily identify CoT reasoning errors .



Case Demo



Thanks for listening !



Looking for PhD position



NTU VIDA Lab



csw0109.github.io

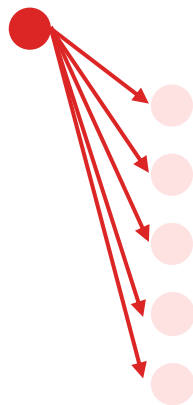
Backup Pages

Appendix: Precision of the diagnose pipeline

ID	count	BIG-Bench		Our		Δ (Our-BIG-Bench)	
		P	R	P	R	ΔP	ΔR
1	91	0.82	0.69	0.75	0.69	-0.07	0.00
2	150	0.36	0.80	0.08	1.00	-0.29	0.20
3	76	0.25	0.73	0.25	0.73	0.00	0.00
4	35	0.27	0.38	0.33	0.63	0.06	0.25
5	59	0.76	0.92	0.79	0.92	0.03	0.00
6	56	0.88	0.47	0.60	0.87	-0.27	0.40
7	70	0.17	1.00	0.06	0.75	-0.11	-0.25
8	148	0.39	0.42	0.24	0.83	-0.15	0.42
9	28	0.40	0.25	0.35	0.63	-0.04	0.38
10	55	0.44	1.00	0.27	0.88	-0.17	-0.12
11	117	0.09	0.33	0.13	1.00	0.04	0.67
12	154	0.56	0.71	0.05	1.00	-0.50	0.29
13	132	0.21	0.86	0.07	0.50	-0.14	-0.36

Table 3: Per-sample comparison of BIG-Bench vs Our Method. Each row corresponds to one evaluation sample, with its statement count and precision/recall. Δ denotes the per-sample difference (Our-BIG-Bench); cell background encodes Δ using a continuous diverging gradient.

1 core error



many flagged steps

- **44.48%** of false positives come from error propagation.

- When a few early core steps are misclassified, many downstream steps inherit the flag.

- In practice the interface mitigates this: a burst of flagged steps traces back to a small set of core steps that can check quickly.



- User prefer AI system with higher recall than precision [1]
- **Future: agreement across multiple verifiers (our pipeline + prompt-based) to cut false positives while preserving recall.**

[1] Kocielnik R, Amershi S, Bennett P N. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems[C]//Proceedings of the 2019 CHI conference on human factors in computing systems. 2019: 1-14.

Appendix: Limitations and future works

Visual clutter at scale

Dense cross-links obscure structure on 100+ step traces.

→ **adaptive granularity: aggregate simple ops into composite nodes.**

The semantic gap

Locating a wrong step is relatively easy; explaining WHY a logic error is wrong is not.

→ **AI-assisted explanations + interactive debugging (edit / fix steps).**

Precision

Many false positives come from propagation of a few core errors.

→ **agreement across multiple verification methods, keeping high recall.**

Generalizability

Validated on verifiable domains (math, logic, retrieval).

→ **softer encodings for creative / legal text: semantic drift, plausibility; agentic pipeline.**

Appendix: Need of Diagnose CoT

- Reasoning is relevant to the result:
High correlation with the reasoning process result with the final output.
- CoT is naturally interpretable:
Easier for user to interpret CoT than internal components
- “Judge a person by the action not the thoughts”